国际微生物大数据平台的应用与启示

刘 柳 马俊才*

中国科学院微生物研究所 北京 100101

摘要 微生物资源与微生物大数据是国家重要战略资源，是人类赖以生存和发展的重要物质基础和生物技术创新的重要源泉，微生物资源及数据的开放共享对微生物资源的开发利用具有重要意义。世界微生物数据中心建设的全球微生物菌种资源目录大数据平台，拥有46个国家的120个国际微生物资源中心的40万株微生物实物资源数据信息。该中心以统一数据门户的形式对全世界科技界和产业界提供微生物菌种资源的信息服务，全面参与国际微生物数据标准制定，并为《名古屋议定书》（*The Nagoya Protocol*）及履约工作在微生物领域的实施提供重要支撑。在此基础上，2017年世界微生物数据中心启动了全球微生物模式菌株基因组和微生物组测序合作计划，从而实现了从微生物资源数据到微生物实体资源共享利用的转变，希望通过微生物大数据平台促进生物大数据产业的发展，进一步推动中国微生物组计划的实施，主导国际微生物组计划，提升我国在微生物领域的话语权。

关键词 微生物资源，微生物大数据，大数据平台，战略资源

DOI 10.16418/j.issn.1000-3045.2018.08.012

微生物资源是人类赖以生存和发展的重要物质基础，是生命科学和生物技术创新的重要源泉。随着生物大数据时代的到来，微生物及其基因资源数据也正呈现爆炸性增长，微生物学研究正从以数据为支撑逐渐向以数据为中心转变，海量数据的整理整合和开放共享对于微生物资源的研究和利用变得至关重要，微生物学已进入了组学数据时代。

与微生物相关的数据资源建设方面，国内许多单位已经分别建立了近百个生物信息资源数据库，数据总量达到PB量级。在国家“863”计划的支持下，我国生物信息技术与平台管理技术体系已经成熟。北京和上海建立了分布式的生命科学基础公共信息分享平台，为国际公共数据库的引进、我国生物学基础科学数据的共享、二次数据库的开发做了大量卓有成效的工作，使我国在

*通讯作者

资助项目：中国科学院战略性先导科技专项（XDA19050301），国家重点研发计划（2016YFB1000605、2016YFC1200801、2016YFC0901702、2017YFC1201202）

修改稿收到日期：2018年8月10日

分布式的生命科学基础公共信息分享平台建设奠定了良好基础。

在大数据的背景下，未来的微生物学研究必将朝着形成一个全方位的微生物资源研究、开发与应用的网络的方向发展，微生物研究各个环节的联系更加紧密，但每个环节的深度也在不断增加，对数据应用必将提出更高的要求。随着云技术的发展，为大规模的数据存储、计算和多样化的分析提供了很好的解决方案。因此，利用云技术，为科学家提供既包括整合型的数据，又能够提供可定制数据分析服务的平台，也将是未来微生物学数据研究的一个重要趋势。

1 微生物资源与微生物大数据是国家重要战略资源

微生物作为最简单的生命体，蕴藏着极为丰富的物种资源和基因资源，微生物丰富的生物多样性也使其成为生物技术和生物产业发展的基石，为人类解决能源、环境危机提供了重要平台[1-4]。目前可培养微生物仅占微生物资源的 1%，而且这些已培养的微生物菌种的利用程度也非常低；因此，微生物资源是一笔巨大的、尚未开发的资源财富，开发和利用微生物资源具有重要的现实意义。微生物资源的有效利用是国民经济可持续发展不可或缺的条件之一，直接影响国家的未来经济发展潜力，也是一个国家重要的战略资源[5]。美国、欧盟、日本等都将微生物资源的开发与利用纳入其战略发展规划，并围绕微生物资源的发掘利用和产业开发等主题进行了中长期的部署。

微生物数据资源是微生物资源共享和开发的关键环节，数据资源的丰富性、准确性和共享水平决定着整个微生物学领域研究和应用的综合能力。与实物资源相比，微生物数据资源是最有可能实现共享的一种资源。通过信息技术，建立统一的数据标准，为微生物资源研究的各个环节提供包括数据管理及共享、数据分析、计算模型等在内的支撑，促进信息资源的共享从而带动微生物资源的开发和利用，对微生物资源研究和生物技术发展具有重要意义。21 世纪初，国际经济合作与发展组织（OECD）推动建设全球生物资源中心网络（Global Biological Resources Centre Network，GBRCN），欧盟也推动过欧洲生物资源及信息共享项目（Common Access to Biological Resources and Information，CABRI），但是这些计划由于缺乏共享机制和技术力量支持等原因，都没能建立一个稳定运行的、成熟的国际性数据平台。

2 微生物资源大数据平台建设

2.1 世界微生物数据中心落户中国

世界微生物数据中心（World Data Center for Microorganism，WDCM）成立于 1966 年，隶属于世界微生物菌种保藏联合会（WFCC）和联合国教科文组织下的全球生物资源中心网络（GBRCN），是全球微生物领域最重要的实物资源数据中心。经过全球竞争，2010 年世界微生物数据中心（WDCM）正式落户于中国科学院微生物研究所。这是落户于我国生命科学领域的第一个世界数据中心，其落户中国标志着我国微生物学研究领域在国际上影响力的大幅提升，也给中国微生物资源研究与利用带来了巨大的发展机遇。迄今，全球共有 76 个国家的 755 个微生物资源保藏中心在 WDCM 注册①。

WDCM 建设和维护了与微生物资源相关的一系列重要数据库，包括全球微生物保藏机构数据库（Culture Collections Information Worldwide，CCINFO）、全球微生物菌种资源目录（Global Catalogue of Microorganism，GCM）、全球微生物参考菌株数据库（Reference Strain Catalogue，RSC）、微生物资源引用数据库（Analyzer of Bioresources citation，ABC）等，是全球微生物领域最重要的实物资源数据平台[6]。

在大数据整合技术研究方面，WDCM 团队开发了生

① Culture Collections Information Worldwide. [2018-07-06]. http://www.wfcc.info/ccinfo/.

物资源引用平台系统，利用先进的数据挖掘手段，从全球超过600万已发表的微生物相关文献、专利、核酸序列和基因组中，进一步提取了微生物资源的后续研究和利用的信息，并开发了参考菌株目录。作为一个跨平台参考目录，该目录整合ISO以及其他国际标准菌种统一编号，推动了全球菌种资源的高标准应用。在数据集成和服务机制上，WDCM团队也进行了积极的探索，使得该平台能够有效地在全球范围进行数据资源的集成，并实现可持续发展。同时，WDCM作为一个合作平台，使我国科学家能够在全球的角度，组织和协调各国的相关力量，建立全球性的合作框架，也让中国有机会逐步在微生物资源的开发应用和数据共享方面占领国际微生物研究前沿和主导地位。截至2018年7月底，平台的累计访问次数已超过20万次。

2.2 倡导全球微生物资源目录合作计划（GCM 1.0），推动微生物数据资源共享利用

为了推动全球微生物数据资源的共享和利用，更好地整合不同来源、不同数据格式的微生物相关的数据，WDCM于2016年9月6日在全球保藏中心之间提出了“全球微生物菌种资源目录国际合作计划”（Global Catalogue of Microorganism），旨在为目前分散在全球各个保藏中心和科学家手中珍贵的微生物资源提供一个全球统一的数据门户。此门户系统覆盖主要保藏中心的重要微生物资源，并且包括微生物资源在采集、鉴定、保藏和应用方面的详细信息。这一国际合作计划建立起了一套统一的全球微生物菌种目录，对主要保藏中心的目录进行标准化整理，提供统一的检索出口。同时，在该目录中集成利用自动化的知识挖掘方法得到的关于微生物资源的文献、专利、序列、基因组等其他知识资源，并开发多种途径的数据检索工具以及数据推送、数据定制服务。

这项计划由中国科学院微生物研究所微生物资源与大数据中心负责具体的信息平台建设、数据标准建立、数据集成与共享实施。目前，已经有来自美国、法国、德国、日本等46个国家的120个国际微生物资源中心正式加入，40万株微生物实物资源的信息汇集到中国团队开发的数据平台②。

2.3 全面参与国际微生物数据标准制定

长期以来，由于各个微生物资源中心采用不同的数据格式进行数据管理和共享，这极大地阻碍了微生物数据交换和在全球范围共享资源的效率。中国科学院微生物研究所微生物资源与大数据中心和WDCM基于其组织的全球微生物菌种目录（GCM）微生物数据资源国际合作计划的工作基础，并经过与国际标准化组织生物技术委员会（ISO/TC 276）及WDCM各国专家的讨论逐渐形成了《微生物资源中心数据管理和数据发布标准（草案）》。经过一年多的筹备，2017年7月，作为ISO/TC276生物样本库与生物资源工作组（WG2）和生物数据处理及整合工作组（WG5）的共同项目，该项目通过了ISO的新工作项目提案（new work item proposal）投票正式立项。该项目现已注册为工作草案（working draft），预计将在2年内正式发布国际标准，并将成为微生物资源数据领域的第一个ISO国际标准。该标准的制定和实施将有助于保证微生物资源数据质量，并提高全球范围微生物数据的兼容性和互操作性，为高效的数据共享和大数据分析提供基础。

2.4 为《名古屋议定书》及履约工作在微生物领域的实施提供重要支撑

我国幅员辽阔，是全球12个生物多样性大国之一，遗传资源极其丰富。但长期以来，我国一直是发达国家获取遗传资源和遗传资源相关传统知识的主要对象，外国机构和个人通过多种非正当手段大量获取我国丰富的生物遗传资源，由此造成的流失数量和价值难以估量，形势十分严峻③。

《生物多样性公约》（*Convention on Biological*

② Global Catalogue of Microorganisms. [2018-07-06]. http://gcm.wfcc.info/.

③ 我国正式加入《名古屋遗传资源议定书》，曾大量流失遗传资源。[2018-07-06]。https://www.thepaper.cn/newsDetail_forward_1524596。

Diversity，CBD）旨在保护濒临灭绝的植物和动物，最大限度地保护地球上的多种多样的生物资源，以造福于当代和子孙后代。我国于1992年6月11日签署该公约，1992年11月7日批准，1993年1月5日交存加入书。2010年10月，联合国《生物多样性条约》第10届缔约方大会（简称“COP10”）通过《名古屋议定书》（*The Nagoya Protocol*，NP）；2014年10月，《名古屋议定书》正式生效。《名古屋议定书》规定通过适当的资金援助和技术合作来保护生物多样性，实现生物遗传资源的可持续利用，其目的在于保障生物遗传资源利益的公平分配。

WDCM大数据平台下的“全球微生物菌种目录系统”（GCM）是一个包含微生物资源的检索、分析和可视化的综合数据库。GCM结合更多的在线目录数据，将菌种资源与核酸序列、蛋白质、参考文献、引文数据等进行关联，并以统一数据门户的形式，对全世界科技界和产业界提供微生物菌种资源的信息服务[7]。GCM对于微生物实物资源从采集、保藏、跨国转移、学术和商业应用以及利益分享的各个环节都能提供有效的数据支持，为《生物多样性公约》及《名古屋议定书》（CBD/NP）在微生物领域的实施和执行提供最重要的支撑。GCM平台及其相关的指导原则，在国际上第一次建成一套完善的可运行的信息平台方案。WDCM在CBD/NP实施方面的相关工作，也符合我国参与CBD工作的主要方向。目前，CBD的信息交换所、国际微生物领域、法律界以及我国环保部专家都对WDCM的相关工作给予了高度认可，对GCM平台对CBD/NP的实施给予了相当的肯定。

3 启动“模式微生物基因组测序、数据挖掘及功能解析全球合作计划”（GCM 2.0）：从微生物资源数据到实物资源的共享利用

模式菌株（type strains）是在给微生物定名、分类记载和发表时，作为分类概念的准则，即以纯菌（可繁殖）状态所保存的菌种。模式菌株由于其参考性和唯一性，对微生物的鉴定、功能研究和大规模组学数据分析都具有重要的价值。目前已知的微生物模式菌株广泛地分布在全球的保藏中心，已测序的微生物基因组还存在大量的空缺。通过对所有已知物种的模式菌株进行组学数据解析，具有重大的科学意义和战略意义。随着测序成本降低和海量数据分析能力的提升，发起大规模的测序计划，开展以序列分析和功能挖掘为基础的研究，已是大势所趋。

2017年10月，由中国科学院微生物研究所牵头，联合全球12个国家，共同发起了“模式微生物基因组测序、数据挖掘及功能解析全球合作计划”（Global Catalogue of Microorganisms 10K Type Strain Sequencing Project）[8]。该计划将在5年内完成超过10 000种的细菌、真菌、古生菌模式菌株基因组测序，覆盖目前已知的全部细菌、古菌模式菌株以及重要的真菌模式菌株，建立全球微生物模式菌株基因组和微生物组测序合作网络，覆盖超过20个国家的30个主要保藏中心，从全球微生物资源保藏中心选择目前未进行测序的模式微生物菌株，完成超过总体90%以上的微生物模式菌株的基因组测序。

作为中国牵头的国际大科学计划，该计划将建立覆盖全球主要合作伙伴，尤其是发展中国家的科技资源共享网，聚集全球微生物领域优势科技资源和顶尖科学人才，帮助解决领域基础和前沿的重大科学问题，也为《生物多样性公约》履约和《名古屋议定书》中的生物资源跨国转移及惠益分享机制等国际合作贡献中国智慧和中国方案，充分体现了我国在微生物领域的科技创新竞争力和国际引领的综合能力。

4 思考与建议

4.1 积极提供数据增值服务，确保全球微生物资源数据合作计划的顺利实施

“全球微生物资源目录合作计划”（GCM 1.0）目前已经有来自美国、法国、德国、日本、中国、印度、越南、巴西等46个国家的120个国际微生物资源中心正式加入，40万株微生物实物资源的信息汇集

到中国团队开发的数据平台。在这120个国际微生物资源中心中，发达国家的和发展中国家的基本上各占一半。这么多发达国家和发展中国家的微生物资源保藏中心之所以向WDCM免费提供他们的数据，是由于其开发的微生物综合大数据平台，能够针对发达国家和发展中国家的保藏中心提供各自所需的个性化增值服务。以欧洲最大的微生物资源保藏中心——德国微生物菌种保藏中心（Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH，DSMZ）为代表，当他们把其保存的3万多微生物资源目录的数据提交到WDCM的数据平台后，我们可以通过开发的数据挖掘系统，清晰地了解全世界科学家和产业界在过去30年里使用DSMZ出售的菌种，撰写了多少科学论文，申请了多少国际专利，产生了多少核酸序列数据，也就是说我们可以告知DSMZ他们在过去30年里为国际学术界做出了多大的贡献。DSMZ所长在撰文描述他们在过去几十年的发展历程时，也使用了我们提供的数据[9]。以越南典型培养物保藏中心（Vietnam Type Culture Collection，VTCC）为代表的发展中国家保藏中心，通过加入WDCM这一全球合作计划，可以方便地使用WDCM的大数据平台，建立自己的对外主页和网上菌种目录数据库；还可以通过WDCM这一全球化信息平台把自己的菌株信息展示给全世界，提高自己的知名度，促进全球微生物资源数据的共享。GCM 1.0是国际微生物资源领域由我国倡导和实施并获得广泛响应的国际合作计划，摸索了一套整合全球微生物资源数据的有效机制，促进了微生物资源数据的全球共享利用，也确立了我国在微生物资源数据共享方面的引领地位。

4.2 利用现有优势，进一步牵头国际微生物模式菌株测序计划，从全球数据共享到实物资源合作

微生物组学也是一个世界各国争相发展的战略性科技领域，美国、日本等发达国家已经部署了支持微生物组研究的国家计划[10]。2016年5月13日美国宣布启动“国家微生物组计划”，相关政府部门携手私营机构投资高达5亿美元，对微生物组进行全面深入的研究，并将研究成果广泛应用于医疗、食品生产及环境保护等重点领域[11]。我们应当以模式微生物基因组测序计划为抓手，依靠我国在微生物资源的研究、测序技术、微生物数据综合分析能力等方面的优势，抓住机遇，尽快启动涵盖人体、农业、环境、传统发酵、新技术等内容的“中国微生物组计划”重点研发专项，并进一步利用该计划建立的国际合作网络，启动中国引领的微生物组国际合作计划，抢占微生物领域的战略制高点。2017年10月，我们在现有全球数据合作的基础上，启动了“模式微生物基因组测序、数据挖掘及功能解析全球合作计划”（GCM 2.0）。目前已经有14个国家的24个保藏中心参加这一计划，并提供相应模式菌株的菌株或者DNA，使我们从早期的全球微生物资源数据共享，进入到实物资源合作阶段。

4.3 通过生物大数据平台推动生物大数据产业发展

BCC Research的报告中指出：“2013年，全球新一代测序和数据分析市场总额为5.1亿美元，至2018年，这一市场总额将增长至76亿美元，复合年增长率达到71.6%”[12]。生物大数据蕴涵着巨大的产业价值，属于国家战略资源。我国是生物多样性和生物技术大国，生物物种、生物资源和生物技术数据极其丰富，这些数据与生物产业息息相关。未来国家的核心竞争力将很大程度上取决于将数据转化为信息和知识的速度与能力。基于大数据的研究和信息发现已经成为生命科学研究新范式和科技创新引擎，并将改变生物产业格局，催生产业新业态。生物大数据平台是科技推动产业发展的桥梁，应该通过政策规划、科研项目布局等多种方式，引导大数据研究成果与产业化应用进行对接，提升企业参与生物大数据研发的积极性，推动我国大数据产业发展。

参考文献

1 Colwell R R. Biodiversity amongst microorganisms and its

relevance. Biodiversity and Conservation, 1992, 1(4): 342-345.

2 Hayat R, Ali S, Amara U, et al. Soil beneficial bacteria and their role in plant growth promotion: a review. Annals of Microbiology, 2010, 60(4): 579-598.

3 Senni K, Pereira J, Gueniche F, et al. Marine polysaccharides: a source of bioactive molecules for cell therapy and tissue engineering. Marine Drugs, 2011, 9(9): 1664-1681.

4 Prakash O, Shouche Y, Jangid K, et al. Microbial cultivation and the role of microbial resource centers in the omics era. Applied Microbiology and Biotechnology, 2013, 97(1): 51-62.

5 段子渊, 黄宏文, 刘杰, 等. 保存国家战略生物资源的科学思考与举措. 中国科学院院刊, 2007, 22(4): 284-291.

6 Wu L H, Sun Q L, Desmeth P, et al. World data centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. Nucleic Acids Research, 2017, 45(D1): D611–D618.

7 Kurtböke I. Microbial Resources: From Functional Existence in Nature to Applications. Cambridge: Academic Press, 2017.

8 Wu L H, McCluskey K, Desmeth P, et al. The global catalogue of microorganisms 10K type strain sequencing project: closing the genomic gaps for the validly published prokaryotic and fungi species. GigaScience, 2018, 7(5): giy026.

9 Overmann J. Significance and future role of microbial resource centers. Systematic and Applied Microbiology, 2015, 38(4): 258-265.

10 刘双江, 施文元, 赵国屏. 中国微生物组计划：机遇与挑战. 中国科学院院刊, 2017, 32(3): 241-250.

11 The White House Office of Science and Technology Policy. FACT SHEET: Announcing the National Microbiome Initiative. Washington: OSTP, 2016.

12 Business Communications Company. Next-generation Sequencing: Emerging Clinical Applications and Global Markets, BIO126C. Wellesley: BCC Research, 2017.

Application and Enlightenment of International Big Data Platform for Microorganism

LIU Liu　MA Juncai*

(Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China)

Abstract　Microorganism resources and microorganism big data are national important strategic resources, material base for human survival and development, and great source of innovation of biotechnology. Opening and sharing of microorganism resources and data are of great significance for development and utilization of microorganism resources. The big data platform of Global Catalogue of Microorganisms (GCM) constructed by World Data Center for Microorganisms (WDCM) records physical resource information data of 400 000 strains of microbes collected by 120 international microbial resource centers in 46 countries, offering microbial strains resource information service to the scientific, technological, and industrial circles all over the world in the form of unified data portal. It fully participates in formulation of international microorganism data standards, and provides important supports for Nagoya Protocol and its implementation in microorganism field. On this basis, WDCM initiated the international microbial genome and microbiome sequencing project in 2017 to realize the transition from microbial resource data to sharing and utilization of physical microbial resources. It is hoped to promote the development of biological big

*Corresponding author

data industry, further drive implementation of Chinese microbiome project, play a leading role in international microbiome project, and raise our voice in the microorganism field through the big data platform for microorganism.

Keywords microorganism resources, microorganism big data, big data platform, strategic resource



刘 柳 中国科学院微生物研究所微生物资源与大数据中心情报信息分析员，理学硕士。主要从事战略情报与学科情报研究。参与了多项来自科技部、中国科学院以及其他国家科研管理部门的项目和课题。E-mail: liuliu@im.ac.cn

LIU Liu Intelligence analyst of The Center for Microbial Resource and Big Data, Institute of Microbiology, Chinese Academy of Sciences (CAS). She received a master's degree from Institute of Botany, CAS. Her main research fields include strategic intelligence and subject intelligence. She participated in many projects sponsored by Ministry of Science and Technology of the People's Republic of China (MOST), CAS, and other national scientific research management departments. E-mail: liuliu@im.ac.cn



马俊才 中国科学院微生物研究所微生物资源与大数据中心主任，世界微生物数据中心主任，世界微生物菌种保藏联合会（WFCC）理事会执委，科技部人类遗传资源管理专家委员会委员。国家“863”计划“微生物数字化信息系统集成关键技术的研发”项目首席科学家。主要研究领域包括：微生物资源和生物技术领域信息化、基于云环境的微生物大数据管理和分析平台。E-mail: ma@im.ac.cn

MA Juncai Received Ph.D. degree from Department of Bio-resources of Mie University in Japan, now he is the director of The Center for Microbial Resource and Big Data, Institute of Microbiology, Chinese Academy of Sciences (CAS). He is also the director of WFCC-MIRCEN World Data Center of Microorganisms (WDCM), and Board Member of World Federation of Culture Collections (WFCC). He is the Principle Investigator of National High Technology Research and Development Program “Key Technology Researches on Microbial Digital Resources Information System”. His research fields cover informalization of microbial and biotechnology resource, cloud based big data management, and analysis system of microbial resources. E-mail: ma@im.ac.cn

■责任编辑：岳凌生